Discussion on: 'Simple or complex statistical models: non-traditional regression models with intuitive interpretations' by Gillian Z. Heller

Peer-reviewed author version

Beyersmann, Jan; Melis, Guadalupe Gomez; Kneib, Thomas; MOLENBERGHS, Geert; Muggeo, Vito; Vansteelandt, Stijn & Heller, Gillian Z. (2024) Discussion on: 'Simple or complex statistical models: non-traditional regression models with intuitive interpretations' by Gillian Z. Heller. In: Statistical modelling, 24 (6) , p. 520 -540.

DOI: 10.1177/1471082X241277642
Handle: http://hdl.handle.net/1942/44662

Discussion on: "Simple or complex statistical models: non-traditional regression models with intuitive interpretations" by Gillian Z. Heller

Jan Beyersmann [1], ORCID: 0000-0002-3793-4611

Guadalupe Gmez Melis [2], ORCID: 0000-0003-4252-4884

Thomas Kneib [3], ORCID: 0000-0003-3390-0972

Geert Molenberghs [4], ORCID: 0000-0002-6453-5448

Vito Muggeo [5], ORCID: 0000-0002-3386-4054

Stijn Vansteelandt [6], ORCID: 0000-0002-4207-8733

[1] Institute of Statistics, Ulm University, Ulm, Germany

[2] Departament dEstadstica i Investigaci Operativa, Universitat Politcnica de Catalunya BarcelonaTech, Barcelona, Spain

[3] Chair of Statistics, Georg-August-Universität Göttingen, Göttingen, Germany

[4] L-BioStat, Medical Faculty, Katholieke Universiteit Leuven, Leuven, Belgium

[5] Dipartimento Scienze Economiche, Aziendali e Statistiche, Universit degli Studi di Palermo, Palermo, Italy

[6] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

Address for correspondence: See the different discussants.

E-mail: amayr@uni-bonn.de.

Phone: (+49) 228 287 16160.

Abstract:

Key words:

1 Comment by Jan Beyersmann

Some words about Gillian's paper - use classical bibtex citations Lesaffre et al. (2009).

1.1 Models for binary data

You can use subsections if you want to.

1.2 Models for time-to-event data

2 Comment by Guadalope Gmez Melis

Text by Lupe.

3 Comment by Thomas Kneib

Text by Thomas.

4 Comment by Geert Molenberghs

A strong merit of the article by Gillian Heller is that it makes us reflect on the models we commonly use for statistical modeling of various data types, in particular binary data and time-to-event data.

Simply said, Heller suggests that we should carefully balance considerations of mathematical elegance and computational convenience on the one hand, and ease of interpretation and communication of results on the other.

The central framework is that of generalized linear models (GLM; McCullagh and Nelder, 1989), built upon the exponential family of distributions. Extensions such as GAMLSS are considered too. The unifying GLM framework led to a tremendous expansion of the modeler's toolkit at the time, which up to then centered largely on the well-developed linear models framework, with analysis of (co)variance and linear regression as its most prominent representatives, as well as ramifications towards multivariate linear regression, MANOVA, and such multivariate techniques as principal components analysis, factor analysis, canonical correlation, and normal-distribution-based discriminant analysis. Prior to the development of GLMs, modeling of, for example, binary and categorical data, proceeded by well-developed but *ad hoc* methods, such as χ^2, Fisher's exact, McNemar, and Cochran-Mantel-Haenszel tests, with (conditional) logistic regression emerging (Breslow et al., 1980).

Heller further points to the important and somewhat unfortunate fact that the modeling of time-to-event data took a different turn, away from GLMs, implying that the close link with other data is under-valued and hence under-used. For example, time-to-event data and count data share the fact that their mean parameters range

over the non-negative half line, making the log link a natural choice.

As logarithms and exponentials map multiplication onto addition and back, the log link enjoys more elegant properties than the various links for intervals (such as the logit and probit links). Still, the log link falls short of the elegance of the identity link, which can be used for data and hence mean parameters on the entire real line.

This leads us to back to models based on normality. They enjoy a large number of properties that do not simultaneously transfer to other data types and the models used in that context. In this sense, while the normal distribution is a specific instance of an exponential family distribution and the linear model is a special case of a GLM, they have a much wider array of convenient properties than is generally the case. Because of the identity link, there is no need to choose between scales that are either mathematically convenient or facilitate interpretation; both apply simultaneously in the normal case. The fact that ordinary linear regression essentially coincides with normal regression makes the model applicable beyond outcome data that are normally distributed. Another useful feature of the normal distribution is the functional independence between the mean and variance(-covariance) parameters, whereas a mean-variance link, at least in part, is present in most GLMs. This link implies that model misspecification can have more severe consequences when non-normal models are used and forces one to deal with over- or under-dispersion in many cases.

Furthermore, the univariate normal distribution has a very natural extension to the multivariate case, while the multivariate normal distribution, in turn, is closed under both marginalization and conditioning.

An issue that should not be overlooked is the potential for misspecification when

choosing between modeling options. When considering three models for the same binary outcome (odds ratio, relative risk, and risk difference), and when there is correction for other covariates, not all three quantities will be simultaneously constant. For example, if a logistic – constant odds ratio – model is used, the relative risk will not be constant. Thus, changing from one scale to the other should be done with caution and upon verification of the fit of the model.

A particular issue is also the fact that the log link for binary data is not range-preserving. Heller addresses this and rightly points to the fact that we have computational tools to enforce valid solutions. Of course, technically it remains possible that certain combinations of covariates, outside of the observed configuration, would lead to non-valid solutions. The large sample behavior with such non-standard links need to be investigated, at least in extensive simulation studies.

While the paper focuses on models for univariate data, there is a variety of extensions towards multivariate, longitudinal, and otherwise hierarchical data that should be kept in mind too. The issues brought forward in this article further deepen in such settings. Also here, the normal case is the fortunate exception. As mentioned by Molenberghs et al. (2013) and Kenward and Molenberghs (2016), the normal distribution is the only one that is self-bridging, apart from the degenerate and Cauchy distribution, and as such the only regular one. That said, the normal distribution is not conjugate to the Bernoulli, Poisson, exponential, or Weibull distributions, implying that the generalized linear mixed model (GLMM; Breslow and Clayton, 1993; Molenberghs and Verbeke, 2005) is less of a natural choice than its common use suggests. Using the conjugate distributions instead (beta for Bernoulli, gamma for the others), while mathematically convenient, comes with its own limitations. Molenberghs et al.

(2010) combined both normal and conjugate random effects into GLMMSs into a single model, in an effort to flexibly model both overdispersion and correlation between repeated measures.

These authors introduced the concept of strong conjugacy, taken to mean that conjugacy between the outcome and conjugate random effects 'survives' the introduction of normal random effects into the linear predictor. Apart from, evidently, the normal case, this property applies to the Poisson model with gamma random effects, and to the Weibull and exponential models also with gamma random effects. It does not apply to the Bernoulli or binomial models with beta random effects. This is one of many peculiarities of the binary case or, more generally, to cases where the mean parameter ranges over a finite interval. This implies that in a GLMM for binary data the fixed-effects parameters are not interpretable as describing the marginal mean function, but rather that of a 'median subject,' i.e., one with all random effects equal to zero. Marginal means can be obtained directly from generalized estimating equations (Liang and Zeger, 1986; Molenberghs and Verbeke, 2005)) or other fully parametric or semi-parametric marginal models.

Related to this, the peculiarity of the binary case also surfaces when three tools are considered to derive marginally interpretable parameters (or functions) from hierarchical models, as reviewed in Molenberghs et al. (2013). The first is to marginalize a GLMM by integrating over the normal random effects in the linear predictor; the second is to consider a hierarchical model built around a marginal mean function (Heagerty, 1999); the third is to replace the normal distribution of the random effects by a so-called bridge distribution that preserves a marginal interpretation of the linear predictor, potentially modulo translation and scaling (Wang and Louis, 2003).

When considering linear models for normally distributed outcomes, the three operations are trivial. For models with a log link (e.g., for counts and time-to-event data) the three operations are relatively straightforward. For example, a marginal mean function from a GLMM with log link has a closed form, and parameters different from intercepts retain their interpretation. In the same case, every sufficiently regular distribution can be used as a random-effects distribution. Even for binary data with probit link, marginalization of the mean leads to a closed form, although the variance components appear in the marginal mean function, and the bridge is formed by the normal distribution, thanks to the distribution's self-bridging (and self-conjugacy) property. However, when the logit link is used, all three operations are different and non-trivial.

It is thus reasonable to add to Heller's comments that even when a distributional and modeling choice is made for convenience in the univariate case, it might not carry over to hierarchical settings.

The fact that time-to-event data took a separate turn, as Heller correctly states, has had consequences for hierarchical models as well, and often the connection between frailty models and random-effects models (the latter in the sense of normal random effects in the linear predictor) is overlooked. Whereas frailties in the sense of gamma random effects at the level of the mean parameter of Weibull models seems a natural choice, care is needed. Molenberghs and Verbeke (2011) showed that the Weibull-gamma frailty model always has only a finite number of finite moments. An extreme example of this, and one to which Heller also refers, is the log-logistic distribution, which does not have any finite moments, a feature frequently overlooked.

In summary, it seems imperative for the statistical modeler to have a good working

knowledge of large classes of models, such as GLM and extensions in the univariate (e.g., GAMLSS) and hierarchical directions (e.g., GLMM and related families), their statistical, mathematical, computational, and interpretation qualities, to facilitate a well-informed, pragmatic choice. While in some cases the interpretation may dominate the choice, in others different considerations may prevail, after which Monte Carlo or approximation calculations are done to arrive at quantities or functions of scientific interest.

5 Comment by Vito Muggeo

Text by Vito.

6 Comment by Stijn Vansteelandt

Text by Stijn.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

This is the place to mention the funding if it is applicable for the paper.

References

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, **88**(421), 9–25.

Breslow, N. E., Day, N. E., and Heseltine, E. (1980). *Statistical methods in cancer research*. International Agency for Research on Cancer, Lyon.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**(3), 688–698.

Kenward, M. G. and Molenberghs, G. (2016). A taxonomy of mixing and outcome distributions based on conjugacy and bridging. *Communications in Statistics-Theory and Methods*, **45**(7), 1953–1968.

Lesaffre, E., Komárek, A., and Jara, A. (2009). The Bayesian approach. In Lesaffre, E., Feine, J., Leroux, B., and Declerck, D., editors, *Statistical and Methodological Aspects of Oral Health Research*, pages 315–338. John Wiley and Sons, Chichester.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**(1), 13–22.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer-Verlag, New York.

Molenberghs, G. and Verbeke, G. (2011). On the weibull-gamma frailty model, its infinite moments, and its connection to generalized log-logistic, logistic, cauchy, and

extreme-value distributions. *Journal of Statistical Planning and Inference*, **141**(2), 861–868.

Molenberghs, G., Verbeke, G., Demétrio, C. G., and Vieira, A. M. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325–347.

Molenberghs, G., Kenward, M. G., Verbeke, G., Efendi, A., and Iddi, S. (2013). On the connections between bridge distributions, marginalized multilevel models, and generalized linear mixed models. *International Journal of Statistics and Probability*, **2**(4), 1–21.

Wang, Z. and Louis, T. A. (2003). Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika*, **90**(4), 765–775.